The Special Case of Scientific Data Sharing with Education

Jillian C. Wallis

Center for Embedded Networked Sensing, UCLA

Stasa Milojevic

Information Studies Dept. Graduate School of Education & Information Science, UCLA

Christine L. Borgman

Information Studies Dept. Graduate School of Education & Information Science, UCLA

William A. Sandoval

Education Dept. Graduate School of Education & Information Science, UCLA

The seemingly simple task of reusing data for science education relies on the presence of scientific data, scientists willing to share, infrastructure to provide access, and mechanisms to share between the two disparate communities of scientists and science students. What makes sharing between scientists and science students a special case of data sharing, is that all of the implicit knowledge attending the data must pass along this same vector. Our work at the Center for Embedded Networked Sensing studying aspects of this data reuse problem has shown us a rough outline of how the future of this data sharing will look. Our approach is to start from the prospective of the scientists, looking for opportunities to support scientific research, and then leveraging the data for reuse by education. The investment needed to capture high quality scientific data necessitates the consideration of reuse by the general population as well as other interested scientific parties.

Introduction

For a digital library to be useful it must fit the needs, activities, and contexts of the people who use it, who create it, and who contribute to it. Here we report on the initial stages of research to develop and deploy a digital library of primary sources in habitat biology for use by scientists, teachers, and high school students associated with the Center for Embedded Networked Sensing (CENS), a large, multi-disciplinary research center. We are studying the information-related practices of these communities as input to the design.

A basic premise of Science and Technology Studies (STS) is that science is a technical practice and a social practice (Star, 1995). In the case of CENS, multiple communities are using the same data collection instruments and resulting data. It is the interaction between technological and social aspects of scientific research that makes designing a system for these communities a complicated problem. When, how, and whether data sharing occurs between scientists is influenced by several conditions, such as whether scientific data exist to be shared, whether those scientists are willing to share those data, and whether mechanisms are available to support sharing. Sharing data among scientists reflects community practices, and these practices are only minimally understood (Hilgartner & Brandt-Rauf, 1994). Even less is understood about the conditions under which scientists will share their data with teachers and their students.

Making scientific data available for use in learning is a means to leverage investments in the technology, the

data, and the associated infrastructure. The CENS community initially consists of researchers across multiple disciplines, teachers, and students affiliated with the Center. As we expand, data generated by our sensor networks will be available to other scientists and to teachers and students in other schools. Sharing data among this diverse array of communities is the long-term goal of the research reported here. CENS offers a rare opportunity to study the generation of scientific data and what conditions affect sharing between these different communities.

CENS as a Testbed

The Center for Embedded Networked Sensing (CENS), at the University of California, Los Angeles (UCLA) (http://www.cens.ucla.edu) is a National Science Foundation Science and Technology Center whose mission is to develop embedded networked sensing systems and to apply these technologies to critical scientific and social applications. The research communities represented in CENS are distinguishable by (i) the kinds of technology and (ii) the kinds of data on which members rely. CENS is focusing on four research areas that could use sensors to augment and in some cases replace existing measuring techniques: contaminant transport, seismic monitoring, marine microorganism observation, and biocomplexity. Contaminant transport monitoring presently has two projects: wastewater recycling and environmental protection application of preventing nitrate impact on groundwater. Seismic sensing and structural monitoring projects continuously record data from the Factor Building on the UCLA campus, the most densely instrumented building in North America, and from a 50-node, ad hoc, wireless multi-hop seismic network. Monitoring of marine microorganisms concentrates on the detection of harmful algae using immuno-based methods. Research on biocomplexity is focused on habitat monitoring, and specifically on developing robust tools that can be operated remotely in both uncontrolled natural settings and agricultural settings. For all of these applications, excepting seismic monitoring, the data collected by sensors forms a backdrop on which biological sampling or the observation of natural phenomena can occur. This contextual data can detect potential points of interest, for instance where the environment exhibits variation, ultimately making the task of hand-collecting data more efficient. CENS is funded for five years (2002-2007) and renewable for an additional five years (2007-2012).

CENS includes a wide array of education-related activities, some research and some outreach. The goal of the educational component of CENS is to develop a science education "pipeline", indoctrinating students into the practices of real science at a young age, and then continuing to develop this interest through increasingly challenging projects and opportunities as they mature. This pipeline is believed to both increase interest in eventually pursuing higher degrees in the sciences and developing higher caliber science students. Education research in CENS focuses on developing inquiry-based science instruction for a diverse grade 7-12 (primary and secondary school) population. Other areas of CENS educational efforts include undergraduate internship opportunities, and graduate fellowships, to complete the pipeline.

Data management, the focus of this paper, is a growing area of interest within CENS. We are beginning to generate large volumes of primary data from sensor networks. These data need to be captured, managed, and maintained in ways that can facilitate use by CENS scientists and by other scientists in their research communities. These same data need to be useful for teachers and students in grades 7-12 for inquiry learning.

Problem Statement

As digital libraries of scientific data grow, it becomes ever more important to understand their use by present and future scientists. If these investments are to be leveraged for educational applications, more needs to be understood about the conditions under which scientists will share their data with teachers and students. An essential component of the required cyberinfrastructure is the provision of tools to support the creation and use of metadata, and that the most effective way of discovering the principles and designing the models on whose basis these tools should be developed is by gaining an understanding of the various metadata-related practices, skills, and requirements of working scientists in multiple communities. We are gathering data on

those practices and requirements with the goal of establishing principles for the design and evaluation of metadata tools. These results will be employed to build and test prototype implementations of those tools. Our aim is to enable scientists to focus on research problems that arise within those domains, and to minimize their concern with data management, metadata application, selection and preservation, and other components of the "information problem."

Making scientific data available for use in learning is just one of the many opportunities that exist for the economic and political leveraging of investments in cyberinfrastructure. Our goal is to understand the information management practices and requirements of the scientists, teachers, and students and to design and deploy digital library services to support all of these communities. While a wide array of projects have involved scientists and scientific data in K-12 instruction, the vast majority of those projects have avoided the dual data management problem by providing students with processed or “canned” selections of data for scripted activities. This approach contributes to our understanding of how inquiry learning can be accomplished but does little to leverage the investment in scientific data production. The ultimate purpose of our project is to enable scientists and students to use the same data, at the same time, in real time.

This multilayered research question must be unpacked into the following series of postulates. There exists some scientific data. There exists some scientific researcher willing to share their scientific data. There exists some architecture to support scientific data sharing. Finally, there exists some means to share scientific data with science students.

Background

Habitat biology

Knowledge about biodiversity and ecosystems is a vast and complex information domain (Schnase, Kama et al., 1997). It is inherently interdisciplinary with roots and applications in a range of different fields. Some of the complexity of the domain arises from the biological complexity of the organisms themselves. The mechanisms used to collect and store biological data are almost as varied as the natural world they document. Data management systems for ecological research have evolved over the last 30 years out of large projects like the International Biosphere Program (IBP) and the US Long-Term Ecological Research (LTER) Program (Brunt, 2000). The role of data management is continuing to change as research projects are becoming broader and more complex. Not only do the data management systems need to support different data types (such as text, numerical measurements, images, sound and video), but they also need to interact successfully with a range of other systems that include geographical, meteorological, geological, chemical, and physical data. Among the biggest challenges to developing useful data management systems is dealing with data diversity (Porter, 2000).

The ecology community has long recognized that data management and metadata are crucial to their work. More emphasis has been placed on standardizing geospatial data, than on non-spatial data. Non-spatial ecological data are more diverse in format and scope. The most important recent developments in ecological data management have been developed by several consortia and are currently overseen by Knowledge Network for Biocomplexity (KNB). The KNB has developed Ecological Metadata Language (EML), an XML-based description standard. EML is a modular exchange standard for communicating metadata. As most XML standards it is almost impossible to use without some front end, and the KNB has developed a Web-interface, called Morpho, which uses a short list of the XML fields to support the data management needs of individual researchers, working with discrete datasets. The front end of Morpho ties back to the KNB's data repository, allowing for the researcher to seamlessly publish their dataset. Another standard is Content Standard for Digital Geospatial Metadata (CSDGM)-Biological Profile. This standard is far less flexible and more monolithic.

Defining scientific data

Defining scientific data is a difficult problem, as the notion of data has different meanings depending on which community is consulted. The experiences and goals of each scientist will affect how data are interpreted (Lynch & Woolgar, 1988). The same dataset may be used and reused on multiple occasions by different communities, in different circumstances, and in different ways. A significant challenge for metadata support is the fact that many of the scientific data currently being generated are infinitely reusable. No a priori method exists of predicting the conditions under which, or for what purposes, any given set of data may be reused.

For example, a habitat biologist may query a sensor network for data on temperature, ambient moisture, and barometric pressure in order to correlate these data with video feeds of bird nests. This same dataset may then be incorporated in a seismologist's study of the potential causes or effects of seismic activity in the area. In turn, the same data could be combined with data collected by a particular temperature sensor, for analysis of the sensor's long-term precision by sensor network engineers. Students could use these data to corroborate data collected locally for a longitudinal study. Yet further afield, the same data could be used by sociologists attempting to see how the sharing of data creates a community, or investigating the effects of data sharing on the quality of scientific publications.

Scientific data are an artifact of the scientific method. As data are massaged by scientists to test hypothesizes, they become facts. They are no longer just sensor readings, they have meaning. This context is very important to interpreting the data. The "stability" (in Lynch & Woolgar (1988) terms) or "fixation" (in Knorr-Cetina (1981) terms) of evidence is established during the social intercourse of peer-reviewed scientific publishing. The capturing of data reveals the experimental design and intentions of the scientist. As these data are changed during the research process, by analysis, validation, being condensed into graphs or figures, the residue of the processes employed by the scientist become apparent. Scientific data evolves through these processes.

Data sharing in the sciences

The forward progress of scientific research relies heavily on the willingness for scientists to share data. The little research that has been done in this area nevertheless strongly indicates that practices often are highly varying and individualistic, and that scientists' motivations both to contribute and to share data are subject to a complex mix of incentives and disincentives (Enyedy, 2003; Enyedy & Goldberg, in press; Sandoval & Reiser, 2003). The extent to which cyberinfrastructure services are adopted by the users for whom they are intended depends at least partly on those potential users' prior evaluation of the short- and long-term benefits that they predict will be the outcome of collaborative activity—activity that may well be perceived as risky or unprofitable, either financially or intellectually. For the working scientist, the provision of access to data may not be as important a personal goal as the retention of control over data. The credit that the scientist receives for initial discovery is a significant disincentive for the sharing of data with others prior to when he or she has finished mining the data and publishing the results. Scientists also worry about the "free rider problem" (Hilgartner & Brandt-Rauf, 1994), where the data-collecting scientist may be unwilling to share if they feel that the other scientist is unlikely to reciprocate (Arzberger, Schroeder et al., 2004; Atkins, 2003; Bowker, 2000a, 2005; Zimmerman, 2006).

Architecture for scientific data sharing

One key component of an integrated framework for data management is a system that provides automated support for the description and annotation of data, so that those data that are wanted or required for a particular purpose are easily identifiable, discoverable, and available in a useful form (see, e.g., (Chervenak, Foster, Kesselman, Salisbury & Tuecke, 2001; Baca, 1998)). Without metadata, data have no context, no meaning, and no potential. When data are separated from the context in which they were generated, there remains no way of reconstructing the relationships between data and context and thus no way of deciding whether any given data are relevant to a given purpose. For data to be meaningful, metadata describing those relationships must be created (whether manually, semi-automatically, or, more typically, automatically) and

preserved (Rajasekar, 2001).

At the minimum, the implementation of a metadata support system will involve the following activities: (i) specification of a standard communication framework (such as is provided by XML) for the communication and exchange of metadata, both among the members of the immediate research community, and between the immediate community and others; (ii) specification of the semantics (meaning) and syntax (structure) of a standard metadata schema (i.e., a standard set of metadata elements), for use by all members of the immediate research community; and (iii) implementation of tools enabling members of multiple communities to supervise the creation (manual, semi-automatic, and automatic), analysis, use, and preservation of metadata.

Scientific data authorship and ownership

Authorship or ownership of data can be difficult to determine, and may vary over the course of a project. While data produced by federal government grants in the US, are by default part of the public domain, control over those data may be distributed across many team members. NSF rules require that data be shared in a reasonable period of time. However, practices vary widely as to how data actually are shared, when, and by whom. The feelings and expression of ownership may be proportional to the amount of effort required of the scientist to collect or clean the data (Pritchard, Carver & Anand, 2004).

Research Methods

Our goal is to understand data practices and functional requirements for CENS ecology and environmental engineering researchers with respect to architecture and policy, and to identify where architecture meets policy. The results reported here were drawn from multiple sources over a three-year period (2002-2005). In the first year (2002-2003), we sat in on team meetings across CENS scientific activities and we inventoried data standards for each area (Shankar, 2003). In year 2 (2003-4), we conducted open-ended interviews with scientists and teams, and continued to inventory metadata standards. We used the results of the first two years to conduct an ethnographic study of habitat biologists. In the current year (2005-6), we are interviewing engineers, scientists, and statisticians about habitat biology data using an interview instrument that came out of the previous year's efforts, and participating in meetings of other CENS groups.

Our population at CENS is comprised of some 70 scientists, mostly faculty, and some post-doctoral researchers, and large and varying body of student researchers. About 30 scientists, post-docs, and engineers are working in the area of habitat biology.

Results

How CENS scientists define data

Scientific data comes in a variety of forms: equations, images, biological samples, computer programs, graphs, etc. are all considered viable sources of data, some of which are easier to store and manage than others. What falls under this umbrella of scientific data is also defined by cultural norms, for instance, a program may be considered data by the computer scientists, and not by the habitat biologists.

Computer scientists in CENS view the measurements taken by the sensors as data, because they draw on these results to assess the viability of the technology: reliability, accuracy, battery life, etc. Habitat biologists are not interested in calibration information, but in measurements that can be verified and cleaned to represent some biological phenomena.

Sharing CENS data

Traditionally habitat biologists work alone or in small teams. Data usually are hand crafted, meaning that the

means devised for data collection are specific to the instrument and project. These scientists often take measurements by hand in addition to the sensor data. These hard-won data are typically stored in Microsoft Excel spreadsheets or modeled in MatLab, which scientists mark up only to the extent necessary for their own understanding of the data. These data products are rarely are reused after the research is published. In the unusual cases when another researcher requests access to a dataset, the scientists may need to mark up the data more thoroughly to show the conditions under which the data were collected.

As mentioned earlier, CENS is bound to NSF's requirements for sharing data, but the urge to make this data available to others runs much deeper. The Center is highly collaboratory in nature, and was designed as such. For every application area of CENS technology there is a multi-disciplinary team composed of the scientific and technology researchers, meant to create an iterative design conversation, constantly improving the technology for a given application area. Shared server space is made available to all of the researchers, to share findings, data, and code. There are frequent inter-group collaborations and equipment exchanges. These researchers have become accustomed to this environment of sharing, and are presented with the right incentives and infrastructure to share data with one another, but the current methods for sharing data between scientists do not scale up to the volume of data that will come off the sensor networks in the near future.

The current state of sharing within CENS has persisted thus far because the technology was still in a state of development, and only recently have equipment deployments collected data that was scientifically interesting. The prototype CENS sensor networks sample regularly, and have accrued a vast set of data. Initially the project aimed to build permanent, autonomous sensor networks generating continuous streams of data. This approach has now been revised to put the human back in the loop. Part of data collection using sensors is to make sure the sensor net is dense enough to capture variations in the variable tested. Fine-tuning this variable requires monitoring by the scientist, at least in the short run. The sensor data also serves as a background for the collection of biological measures, such as water samples. Adding this important data to the sensor data would be ideal, but is more difficult to implement because of the possible ownership issues on the part of the collecting researcher.

The original idea of tapping into data streams minimized concerns about data ownership. Once established for an on-going project by a team of scientists, these data could be made available for teaching applications concurrently (provided the activities did not interfere with the scientific projects). Project-based sensor networks that may be short term offer fewer opportunities to provide data directly to teaching applications.

Architecture models

In parallel with studying the science and engineering research teams, we have continued to work with the education team developing the k-12 inquiry learning modules to identify data requirements. As our understanding of the needs of each group grew, we postulated three scenarios for supporting and bridging the communities: common metadata models, packaged learning objects, and filters and tools.

Scenario 1: common metadata models

The habitat biology community is beginning to converge around the Ecological Metadata Language (EML) (Ecological Metadata Language (EML), 2004; Borgman, Leazer et al., 2004). EML describes spatio-temporal variables more thoroughly than does the current CENS schema and provides a means to share data across research projects. However, EML is optimized for describing data, and not the derivation of data (i.e., sensor networks). EML is deficient for describing sensor data, since few ecologists currently use sensor-derived data. The emerging SensorML is a modeling language for describing resources for sensor management and discovery but it does not describe sensor-derived data itself. Both EML and SensorML are XML-based standards and both are extensible. The CENS scientists at the James Reserve have determined that a combination of these two formats will serve their local needs and will assist them in sharing data with the larger biocomplexity

community. The biocomplexity community long has recognized that data management and metadata are crucial to their work. The lack of standards, diversity of data formats, and the lack of well-documented datasets have slowed cross-institutional and longitudinal research in ecology. Although large datasets are available, they are not described in a consistent way so that researchers can search for patterns over time. A primary concern in this area is standardizing geospatial data (Michener, 1997, 1998; Michener & Brunt, 2000).

In contrast to the concern of scientific metadata models for describing data, the widely-used education metadata models (LOM, GEM, SCORM) (LOM (Learning Object Metadata); SCORM (Sharable Content Object Reference Model). D'Avolio, Borgman et al., 2004) are concerned with describing scripted activities in which data can be used, but do not provide data elements for describing the scientific data themselves. Rather, these models include elements for the grade level, educational objectives, equipment requirements, class time requirements, and so on. They were developed to describe static, primarily text-based learning objects, and not dynamic datasets.

In the early stages of our exploration, the most obvious solution appeared to be a common metadata model that would meet the needs of the scientists and of the educational applications. We devoted many months to exploring the available metadata models in each of these domains, analyzing each in comparison to identified needs. Metadata includes information necessary to understand and effectively use data, such as documentation of the dataset contents, its context, quality, structure, and accessibility. The choice of metadata format is an important technical and economic matter. If we could identify one format or a combination of formats that would serve the needs of all concerned, the development and deployment of digital library services for CENS would be simplified greatly.

However, as has been reported in other research on the use of digital libraries, metadata choices are also epistemic choices, since metadata models represent the tacit knowledge and epistemological perspectives of the communities that create and use them (Bowker, 2000a; Bowker, 2000b; 2000c; Van House, 2003). This is also the case in CENS communities. Metadata models for scientific applications and metadata models for educational applications serve very different purposes. Much to our dismay, we found that the available models in wide use for science and for education are fundamentally incompatible.

Metadata models in use by the CENS habitat biology researchers and by others in the habitat biology community describe the data (e.g., time, date, sensor location), while educational metadata models describe the educational activity (e.g., grade, level, resources required for the activity, time to perform the activity, educational standards, etc.). Our survey of the available metadata standards made clear that there is no overlap in data elements between the metadata formats currently in use by the scientific and educational communities we are studying. The fundamental problem in reconciling metadata standards for scientific data and educational applications is that these two standards were developed to serve different purposes.

While metadata schemas are in theory bridging techniques (Marshall, 2003), we found that they cannot bridge the chasm between scientific and educational applications. We tried to create crosswalks from the existing metadata schemas to the CENS schema, hoping to find one that would cover most of our needs. Lacking any intersection between scientific metadata models of interest to our scientists and those available for educational applications, we abandoned this approach, at least for the time being.

Scenario 2: packaged learning objects

The next plan we considered was to follow the route of most other science education projects, which is to create independent learning objects that could be described with educational metadata models. We considered this approach only briefly, however, as it would constrain us to the use of archival data and scripted scientific activities. We would have lost the essential advantage of the CENS approach, which is to use real-time data, to allow open-ended inquiry and hypothesis generation, and to provide students with the ability

to conduct experiments using remote scientific instruments.

Scenario 3: filters and tools

After rejecting the common metadata and packaged learning objects approaches, we pursued a third direction. The new direction follows from our commitment to high quality scientific data being paramount to CENS' mission, and a confirmation that our first priority must be to make these data useful and usable to scientists. If scientists cannot rely on our data management methods for their own work, the data will be of little value for educational applications.

Our current approach has two components. First, as noted earlier, we are working with the James Reserve team to assist them in developing systems and methods to manage their data using the Environmental Markup Language (EML) and SensorML. In this way, they are assured that their data will remain useful for local purposes and for sharing within the larger habitat biology community. Second, we are working closely with the education team developing the inquiry learning modules to assist them in building filters and tools that will make the scientific data useful to teachers and students (grades 7-12). The pilot module, field tested in spring, 2004, and deployed in fall, 2004, addresses "interdependence in nature," and follows California educational standards. Specific learning activities include analyzing correlations between weather and plant adaptations, such as leaf size.

Lacking the domain knowledge, experience, and data analysis skills of scientists, students cannot conduct their own studies of interdependence in nature without some assistance. The tools and filters will provide assistance by simplifying the scientific tasks and the amount of data available, thus removing some of the "messiness" of real science. Tools and filters for the initial modules will provide a considerable amount of assistance, usually known as "scaffolding." As students become more skilled at scientific processes and data interpretation, subsequent modules will provide less scaffolding, gradually moving students toward the tools, interfaces, and full datasets available to scientists. For example, in the pilot modules, the student interface will provide access to only a few variables (e.g., location, temperature, time period, light measurement) to answer a limited number of questions dealing with leaf adaptation to different microclimate conditions. The data will be available to students at a lower granularity than is available to the scientists. For example, temperature measurements might be taken every minute, but students could only graph hourly variations. Variables were selected for the initial pilot that would show sufficient differences between leaves at different locations that the relationships between factors would be apparent. Thus students are working with real data, but the data have been filtered to a selected subset and the tools provide only a few analytical capabilities. Later sets of tools and filters will offer a larger variety of data elements and will enable queries on both animals and plants. Thus students will gradually gain the scientific knowledge and data analysis skills that move them closer to becoming scientists, with the ultimate goal that they will be capable of conducting their own research with the native data interfaces designed for scientists.

Sharing scientific data with students

During the initial stages of the development of inquiry learning modules for science students, there was only one available installed sensor network at James Reserve. While this network was pulling in data, it was not pulling in any interesting data to scientists and especially not to science students. In effect this technology was not mature enough to bring into the classroom or the laboratory. In place of using this sensor networks, we installed a small sensor network in the Santa Monica Mountains. The network was composed of three main sites at three different altitudes, each with sensors to measure temperature, humidity, and barometric pressure. This generated real data, but not necessarily data that any scientist would be interested in using. The data was at least interesting to science students because the varied altitudes demonstrated the effects of heat and moisture on leaf shape. This data was also unencumbered by the sharing and intellectual property issues discussed above.

The students required an interface to structure their experience with the data. The inquiry module formed this structure by devising a guiding question around which their activities coalesced. In the case of the Plant Module, the students were learning about the compromise between photosynthesis and transpiration within the plant. All of these smaller activities lead to their culminating answer about how plants adapt to their surroundings. Plants are selected for leaf area that maximizes photosynthesis, but minimizes water loss due to evaporation. This selection is affected by altitude, temperature, and humidity. The students answer this bigger question and are then advised to use the data to support their statement. The figure below shows the interface given to students participating in the module activities. It consists of a visual topography for the locations of the three sensor stations, images of leaves around each weather station, as well as a graphing tool that simplifies the interaction between the students and the data.

In this first module then, the students were using canned data. This was successful in that the students learned about photosynthesis and transpiration, but they were not answering new questions. If they had been given access to a wider array of data within this structural interface, or scaffolding, real data that scientists also had access to, they could be asking new questions. In the second module that is currently being designed, the students will be learning about cellular respiration through observation of bacteria that contribute to contaminant transport. While the notion of cellular respiration is not in dispute, what role these bacteria play in the production of methylated mercury is not entirely known. These students will be contributing to the scientific body of knowledge, and experience that will positively impact their experience and in turn their comprehension of science.

Discussion

Making scientific data useful for teaching high school science while maintaining it in a form useful to scientists is a much harder problem than it may appear, and one that has received little research attention. One reason for the difficulty is that scientists and students collect and analyze data for different purposes. The second reason is that scientists, teachers, and students bring far different skill sets and epistemologies of science to the use of scientific data. Scientists have established discipline-specific practices to select, collect, organize, analyze, store and disseminate data. These practices reflect a tacit understanding about what the nature of science is, what reasonable questions are, what knowledge claims should look like, and what sorts of evidence are expected to support such claims. Primary and secondary school teachers and students generally lack deep subject knowledge, research methods expertise, and knowledge of data management practices. Thus to achieve the leverage of scientific data for educational use and to maintain the systems' value to the research community, we need to manage scientific data in ways that are useful and usable for communities with very different goals and great disparity in domain knowledge and data management skills. Through inquiry learning using real scientific data in real time, we want to bring students to science and not vice versa.

Digital libraries are complex systems that support many activities associated with the seeking, use, creation, and sharing of information. These activities are embedded in community practices that may vary widely from one community to another. Thus for one digital library to serve multiple communities, design must be based on an understanding of practices in each of the communities and on the relationship between those practices. In our research with scientists, teachers, and students associated with CENS, we have found that these communities differ in ways that are critical to the design of digital libraries. They have very different levels of knowledge about the scientific domain, and about the use and analysis of scientific data. Yet our goal in this project is to bring them closer together by sharing access to primary scientific data being produced by CENS research projects. We wish to facilitate inquiry learning by students in grades 7-12 by providing them with access to real scientific data, in real time, and with tools and services to make use of those data.

Figure: Screenshot of plant module student interface to data and contextual information.

Conclusion

What we are seeing with the CENS community is a great willingness to continue to mature application areas, such as habitat biology, and to cultivate an atmosphere of sharing. In order to reach audiences beyond the confines of the Center, new infrastructure will need to be developed to support the scientific data lifecycle. This lack is a significant barrier not only to the initial generation, discovery, and selection of data, but also to the subsequent reuse of the same data by multiple communities of scientists and nonscientists. An essential component of the cyberinfrastructure required to advance science is the provision of tools to support the creation and use of metadata, without which the data merely form a meaningless string of bits. The most effective models and tools are those based on an understanding of the various metadata-related practices, skills, and requirements of working scientists in multiple communities.

Furthermore, communities other than domain scientists may benefit from strategies to improve the level and quality of access to scientific data. These include science educators, science learners, science policymakers, science activists, and a huge population of science-oriented laypeople. The existence of a sizeable cadre of knowledgeable nonscientists (especially scientists-to-be) is an important predictor of future scientific progress. By involving students in inquiry-based learning-essentially, a means of learning science by doing it-science educators will nurture future generations of scientists who have the knowledge and skills required to succeed in the highly collaborative, highly data intensive world of e-science. An essential prerequisite for effective inquiry-based learning is the provision of access to data of the same kinds that scientists use in practice-data that, ideally, are generated in real-time, and that students can manipulate in the same kinds of ways in which they are filtered, organized, and visualized by scientists.

In addressing our original research problem, we have found that there exists some scientific data worth reuse being collected by CENS instruments. We found the developing culture of collaboration and sharing being nurtured by CENS. We found the beginnings of infrastructure, the emerging data structures and standards, to make sharing of data possible. We also found mechanisms, the use of filters and tools, to repurpose data for k-12 science students. All of these pieces to the sharing with education puzzle are still in the beginning stages, and it will be a number of years before they are completely formed. In the meantime, we will continue to study

these issues and assist in the development of infrastructure and policy.

Acknowledgements

Also contributing to the educational research are Noel Enyedy and Jonathan Furner, both of the Graduate School of Education & Information Studies, UCLA. CENS is funded by National Science Foundation Cooperative Agreement #CCR-0120778, Deborah L. Estrin, UCLA, Principal Investigator. CENSEI is funded by National Science Foundation grant #ESI-0352572, W. A. Sandoval and C. B. Borgman, Principle Investigators.

References

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., Moorman, D., Uhlir, P. & Wouters, P. (2004) An International Framework to Promote Access to Data *Science* 303(5665): 1777-1778

Atkins, D. E., et al (2003) *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon panel on Cyberinfrastructure* National Science Foundation http://www.communitytechnology.org/nsf_ci_report/

http://www.cise.nsf.gov/evnt/reports/toc.htm Visited: September 26, 2003

Baca, M. e. (1998) *Introduction to Metadata: Pathways to Digital Information* Los Angeles, Getty Information Institute

Borgman, C. L., Leazer, G. H., Gilliland-Swetland, A. J., Millwood, K. A., Champeny, L., Finley, J. R. & Smart, L. J. (2004) How Geography Professors Select Materials for Classroom Lectures: Implications for the Design of Digital Libraries *Joint Conference on Digital Libraries, Tucson, AZ* Association for Computing Machinery

Bowker, G. C. (2000a) Biodiversity datadiversity *Social Studies of Science* 30(5): 643-683

Bowker, G. C. (2000b) Mapping biodiversity *International Journal of Geographical Information Science* 14(8): 739-754

Bowker, G. C. (2000c) Work and information practices in the sciences of biodiversity *VLDB 2000, Proceedings of 26th international conference on very large data bases, Cairo, Egypt* Kaufmann

Bowker, G. C. (2005) *Memory Practices in the Sciences* Cambridge, MA, MIT Press

Brunt, J. W. (2000) Data management principles, implementation and administration *Ecological data: Design, management and processing* Michener, W. K. & Brunt, J. W. Oxford, Blackwell Science: 25-47

Chervenak, A., Foster, I., Kesselman, C., Salisbury, C. & Tuecke, S. (2001) The data grid: Towards an architecture for the distributed management and analysis of large scientific datasets *Journal of Network and Computer Applications* 23: 187-200

D'Avolio, L., Borgman, C. L., Champeny, L., Leazer, G. H., Gilliland-Swetland, A. J. & Millwood, K. A. (2004) *From Prototype to Deployable System: Framing the Adoption of Digital Library Services* Alexandria Digital Earth Prototype Project, UCLA: 11 pages

Ecological Metadata Language (EML). (2004) http://knb.ecoinformatics.org/software/eml/ Visited: 25 November 2004

Enyedy, N. (2003) Knowledge construction and collective practice: At the intersection of learning, talk, and social configurations in a computer-mediated mathematics classroom *Journal of the Learning Sciences* 12(3): 361-408

Enyedy, N. & Goldberg, J. (in press) Inquiry in GLOBE classrooms: Developing classroom communities for understanding through social interaction *Journal for Research in Science Teaching*

Hilgartner, S. & Brandt-Rauf, S. I. (1994) Data Access, Ownership and Control: Toward Empirical Studies of Access Practices *Knowledge* 15: 355-372

Knorr-Cetina, K. (1981) *The manufacture of knowledge: An essay on the constructivist and contextual nature of science* Oxford, Pergamon Press

LOM (Learning Object Metadata) http://ltsc.ieee.org/wg12/

Lynch, M. & Woolgar, S. (1988) Introduction: Sociological orientations to representational practice in science *Representation in Scientific Practice* Cambridge, MA, MIT Press: 1-19

Marshall, C. C. (2003) Finding the boundaries of the library without walls *Digital library use: Social practice in design and evaluation* Buttenfield, B. P. Cambridge, MA, MIT Press: 43-64

Michener, W. K. & Brunt, J. W., Eds. (2000) *Ecological Data: Design, Management and Processing* Oxford, Blackwell Science

Michener, W. K., J.H. Porter, & S.G. Stafford (Eds.) (1998) *Data and information management in the ecological sciences: a resource guide* Albuquerque, NM, LTER Network Office, University of New Mexico

Michener, W. K., J.W. Brunt, J. J. Helly, T.B. Kirchner, & S.G. Stafford (1997) Nongeospatial metadata for the ecological sciences *Ecological Applications* 7(1): 330-342

Porter, J. H. (2000) Scientific databases *Ecological data: Design, management and processing* Michener, M. K. & Brunt, J. W. Oxford, Blackwell Science: 48-69

Pritchard, S. M., Carver, L. & Anand, S. (2004) *Collaboration for knowledge management and campus informatics* http://www.library.ucsb.edu/informatics/informatics/documents/UCSB_Campus_Informatics_Project_Report.pdf Visited 14 November 2005

Rajasekar, A. a. R. M. (2001) Data and metadata collections for scientific applications *High-Performance Computing and Networking. 9th International Conference, HPCN Europe 2001. Proceedings (Lecture Notes in Computer Science Vol.2110). Springer-Verlag. 2001, Berlin, Germany* 2110: 72-80

Sandoval, W. A. & Reiser, B. J. (2003) Explanation-driven inquiry: integrating conceptual and epistemic supports for science inquiry *Science Education* 87: 1-29

Schnase, J. L., Kama, D. L., Tomlinson, K. L., Sanches, J. A., Cunnius, E. L. & Morin, N. R. (1997) The Flora of North America digital library: A case study in biodiversity database publishing *Journal of Network and Computer Applications* 20: 87-103

SCORM (Sharable Content Object Reference Model) http://www.adlnet.org/index.cfm?fuseaction=scormabt

SensorML (Sensor Modeling Language) http://vast.uah.edu/SensorML/

Shankar, K. (2003) Scientific data archiving: the state of the art in information, data, and metadata management Retrieved September 26, 2003 from http://cens.ucla.edu/Education/index.html

Star, S. L. (1995) The politics of formal representations: Wizards, gurus and organizational complexity *Ecologies of Knowledge: Work and Politics in Science and Technology* Albany, NY, State University of New

York Press

Van House, N. A. (2003) Digital libraries and collaborative knowledge construction *Digital library use: Social practice in design and evaluation* Bishop, A. P., Van House, N. & Buttenfield, B. P. Cambridge, MA, MIT Press: 271-296

Zimmerman, A. (2006) New Knowledge from Old Data: The role of standards in the sharing and reuse of ecological data *Science, Technology, and Human Values*